Assignment2 - Step 1

Phoebe Zhou

September 2025

Dataset Setup and Exploration Report

For this first step I worked with the RAG Mini Wikipedia dataset, which provides two complementary subsets: a set of question–answer pairs (918 items) and a collection of Wikipedia passages (around 3,200 entries). I accessed both through Hugging Face, verified the schema, and exported them into the project's standard format (gold.jsonl for Q/A and corpus.jsonl for passages). This ensures the files can be plugged directly into the later RAG pipelines.

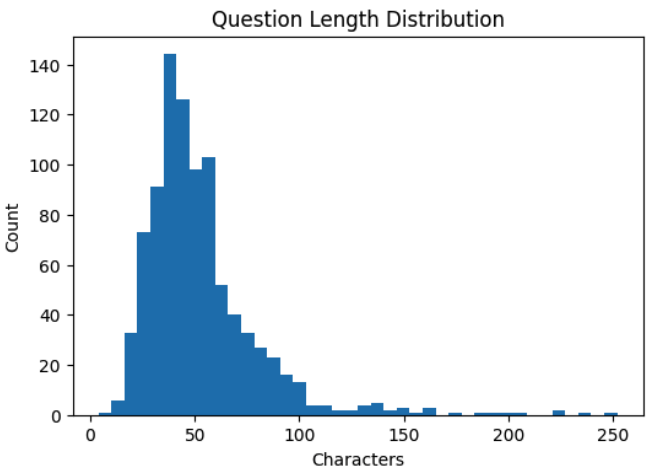
1. Understanding the data.

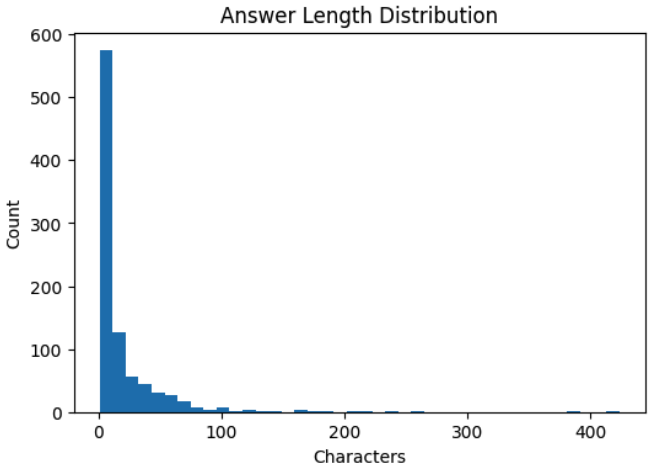
Each record has an id, a natural-language question, and its answer. Most questions are short, averaging about 50 characters, and the answers are even shorter, typically just a phrase or a single word. This indicates the fact-checking style of the dataset. For example, many of the questions are simple factual prompts like "Was Abraham Lincoln the sixteenth President of the United States?". And the answers don't require long explanations, just the correct piece of information.

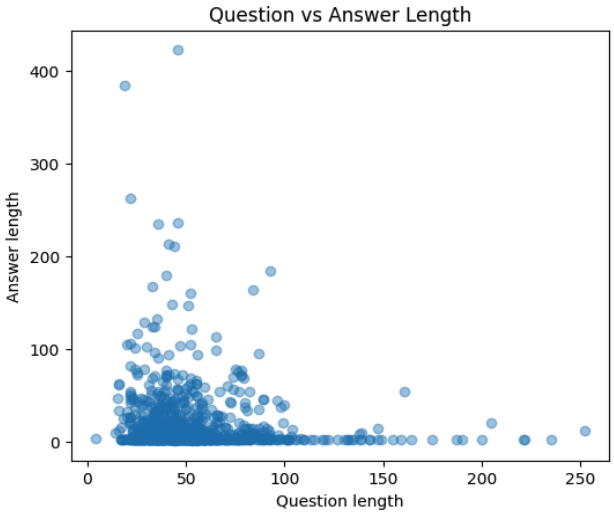
The passages are much longer, with an average length close to 400 characters but a wide range. Some are only a handful of characters, while others stretch into multiple paragraphs. That variability will matter when tuning retrieval, since very short passages may not provide enough context and very long ones may need to be chunked.

2. from data exploration.

Looking at distributions confirmed the intuition: questions are relatively consistent in length, while answers cluster tightly at the short end. A scatter plot of question length vs. answer length showed no real correlation: short questions can have surprisingly long answers and vice versa. The passages show a heavy long-tail distribution, with a small fraction dominating in length. I also noticed that passages lack titles, which slightly reduces contextual richness compared to full Wikipedia articles.


Question Length Distribution
Count
Characters
0
20
40
60
80
100
120
140
0
50
100
150
200
250


Answer Length Distribution
Count
Characters
0
100
200
300
400
500
600
0
100
200
300
400


Question vs Answer Length
Answer length
Question length
0
100
200
300
400
0
50
100
150
200
250